

**Manuscript version: Author's Accepted Manuscript**

The version presented in WRAP is the author's accepted manuscript and may differ from the published version or Version of Record.

**Persistent WRAP URL:**

<http://wrap.warwick.ac.uk/129156>

**How to cite:**

Please refer to published version for the most recent bibliographic citation information. If a published version is known of, the repository item page linked to above, will contain details on accessing it.

**Copyright and reuse:**

The Warwick Research Archive Portal (WRAP) makes this work by researchers of the University of Warwick available open access under the following conditions.

Copyright © and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable the material made available in WRAP has been checked for eligibility before being made available.

Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

**Publisher's statement:**

Please refer to the repository item page, publisher's statement section, for further information.

For more information, please contact the WRAP Team at: [wrap@warwick.ac.uk](mailto:wrap@warwick.ac.uk).

# MILAMP: Multiple Instance Prediction of Amyloid Proteins

Farzeen Munir, Sadaf Gul, Amina Asif and Fayyaz-ul-Amir Afsar Minhas\*

**Abstract**— Amyloid proteins are implicated in several diseases such as Parkinson's, Alzheimer's, prion diseases, etc. In order to characterize the amyloidogenicity of a given protein, it is important to locate the amyloid forming hotspot regions within the protein as well as to analyze the effects of mutations on these proteins. The biochemical and biological assays used for this purpose can be facilitated by computational means. This paper presents a machine learning method that can predict hotspot amyloidogenic regions within proteins and characterize changes in their amyloidogenicity due to point mutations. The proposed method called MILAMP (Multiple Instance Learning of Amyloid Proteins) achieves high accuracy for identification of amyloid proteins, hotspot localization and prediction of mutation effects on amyloidogenicity by integrating heterogeneous data sources and exploiting common predictive patterns across these tasks through multiple instance learning. The paper presents comprehensive benchmarking experiments to test the predictive performance of MILAMP in comparison to previously published state of the art techniques for amyloid prediction. The python code for the implementation and webserver for MILAMP is available at the URL: <http://faculty.pieas.edu.pk/fayyaz/software.html#MILAMP>.

**Index Terms**—Amyloid, Amyloidogenic Hotspots, Amyloid Prediction, Multiple Instance Learning, Machine Learning.

## 1 INTRODUCTION

Amyloids are formed when many copies of certain polypeptide chains stack together or aggregate in a cross  $\beta$  formation as fibres [1][2][3]. A number of different biological mechanisms are involved in amyloidogenesis such as mutations, errors in protein synthesis, intrinsic protein disorder, environmental conditions, maturation and proteolysis, etc. [4]. Amyloid fibers are insoluble, resistant to the action of protease and are implicated in at least 50 known disease in humans such as Parkinson's, Huntington's, Alzheimer's, Diabetes mellitus type 2, prion diseases, etc. [5][6]. The literature points out that specific subsequences, called hotspots, in amyloid proteins can act as seeds for amyloidogenesis [7][8][9]. The identification of these regions can help biologists understand the biological function of such proteins. Similarly, the study of effects of point mutations on amyloid forming proteins or peptides is interesting from a biological perspective [8]. Given the crucial role played by amyloid proteins in many diseases and their interesting biochemical properties, correct identification and prediction of amyloid proteins, their hotspot regions and the effect of mutations on their amyloidogenicity is very important.

Determination of amyloid forming proteins through biological and biochemical assays is time consuming and expensive [10]. Bioinformatics methods can be used to improve the throughput of these experiments. However, computational prediction of amyloid forming proteins, their specific hotspot regions responsible for aggregation and the effects of point mutations on amyloidogenicity are challenging problems because amyloid forming proteins share little sequence or structural similarity [9][11][12]. Several computational amyloid prediction methods exist in the literature. Broadly, these methods can be divided into structure and sequence-based methods. Structure based methods, such as Aggrescan 3D [13] and AggScore [14], utilize the 3D tertiary structure of the protein in their prediction but are limited by the constraint that the tertiary structure of the protein must be available for testing. This constraint can become a limitation in the applicability of these techniques as obtaining 3D protein structures is time consuming and expensive. Prediction of protein structure through computational methods can also become a computational bottleneck in large-scale screening of candidate amyloid-forming proteins. Consequently, sequence-based methods are more widely used. AGGRESKAN [15][16], FoldAmyloid [17], and Pawar et. al. [18] [19] predict amyloidogenesis by relying on aggregation propensities of an individual residues in a polypeptide chain. Zyggregator [20] and TANGO [21] use individual residue aggregation propensities and  $\beta$ -structural conformation properties to predict amyloid forming proteins. Waltz uses the information from position specific scoring matrices (PSSMs), physicochemical properties of amino acids and structure derived from FoldX program to score hexapeptides based on their amyloidogenicity [22] [23]. APPNN is a neural network based amyloid predictor that uses biochemical and physicochemical properties of amino acids for prediction [24].

- Farzeen Munir is with the Gwangju Institute of Science and Technology, Korea, C 305, 123-cheomdangawagi-ro oryong dong, Gwangju 61005 South Korea. E-mail: [farzeen.munir@gist.ac.kr](mailto:farzeen.munir@gist.ac.kr).
- Sadaf Gull is with the PIEAS Biomedical Informatics Lab, Department of Computer & Information Sciences, Pakistan Institute of Engineering and Applied Sciences, PO Nilore, Islamabad 45650, Pakistan. E-mail: [sadafzakarkhan@gmail.com](mailto:sadafzakarkhan@gmail.com).
- Amina Asif is with the PIEAS Biomedical Informatics Lab, Department of Computer & Information Sciences, Pakistan Institute of Engineering and Applied Sciences, PO Nilore, Islamabad 45650, Pakistan. E-mail: [a.asif.shah01@gmail.com](mailto:a.asif.shah01@gmail.com).
- Fayyaz Minhas is with the PIEAS Biomedical Informatics Lab, Department of Computer & Information Sciences, Pakistan Institute of Engineering and Applied Sciences, PO Nilore, Islamabad 45650, Pakistan. E-mail: [af-sar@pieas.edu.pk](mailto:af-sar@pieas.edu.pk), [fayyazaf-sar@gmail.com](mailto:fayyazaf-sar@gmail.com).

Burdukiewicz et al. use sequence n-gram analysis for identifying amyloidogenic motifs [25]. Similarly, the *FISH Amyloid* method uses amino acid co-occurrence for predicting amyloid segments in proteins with an AUC-ROC score of 80% to 95% for experimental and computationally generated datasets [26]. MetAmyl [27], AmylPred [28] and AmylPred2 [29] are ensembles of already existing amyloid predictors which give better prediction accuracy than their constituent methods. A number of sequence databases of amyloid sequences, such as WALTZ-DB [30], AmyLoad [31], etc. are available for development of machine learning based methods for amyloid prediction. Databases that contain information about the role of amyloid proteins in different diseases are also available [32]–[34].

It is important to note that no computational method exists that can simultaneously predict whether a protein will form an amyloid or not, locate its hotspot regions and analyze the effect of mutations on amyloidogenicity. Such a multi-task predictor can exploit common predictive patterns across all three tasks for producing accurate predictions. Although, these prediction problems are related to each other, conventional machine learning techniques do not allow the development of a single model for making predictions for all three tasks as the nature of annotated data available for them varies greatly.

In this paper, we have proposed a novel machine learning approach called *Multiple Instance Learning for AMyloid Prediction* (MILAMP) for prediction of amyloid proteins, their hotspots regions as well as changes in aggregation propensities due to point mutations. Modeling of these three problems through multiple instance learning allows us to leverage similarities among the three prediction tasks and use existing annotated data more effectively to produce a unified and highly accurate predictor. We have observed similar improvements in prediction accuracy of prion proteins through multiple instance learning in our previous work [35]. We have evaluated the performance of the proposed amyloid prediction method on several datasets. We have also performed a large-scale data analysis of amyloid formation propensity of various proteins in the protein data bank (PDB). In addition to that, we also report the performance of the proposed method on recently published amyloid peptides that were not part of our training set.

## 2 MATERIALS AND METHODS

### 2.1 Datasets

For the development of MILAMP, the following datasets have been used. It is important to note the differences in structures and annotations across these datasets to better understand our motivation for applying multiple instance learning to this problem.

#### Dataset-1 (DS1)

DS1 consists of a total of 304 hexapeptides collected from various sources by Familia et. al. [24] with 168 (positive) peptides experimentally verified to form amyloid fibers *in vitro* and 136 (negative) non-amyloid peptides. Detailed information about this dataset is available in supplementary file. This dataset has been used as a training set in the development of the proposed scheme.

#### Dataset-2 (DS2)

DS2 has also been taken from the work by Familia et. al. [24] and it consists of 483 proteins with different polypeptide chain lengths. It contains 341 amyloid proteins whereas the remaining 142 proteins do not exhibit amyloid formation *in vitro*. A *CD-Hit-2D* [36] sequence similarity clustering analysis of sequences in the largest curated amyloid sequence database (*AmyLoad*) with DS2 reveals that *AmyLoad* contains only 3 novel sequences with less than 40% sequence identity to DS2. All other sequences in *AmyLoad* share more than 40% sequence identity with sequences in DS2 [31]. Therefore, we have chosen to use this dataset instead of *AmyLoad* as it allows us direct comparison with previous papers as well. It is important to note that amyloid-forming hotspots have not been annotated for these proteins. This dataset has been used in cross-validation based performance assessment of MILAMP as discussed in section 2.4. Detailed information about the dataset is available in the supplementary file.

#### Dataset-3 (DS3)

DS3 consists of 33 proteins from amyloids that have been used for evaluation in MetAmyl [27]. For each protein in this dataset, the hotspot regions responsible for fibril formation are annotated. Altogether, a total of 70 experimentally validated amyloid-forming hotspot regions are marked. This dataset is used for testing the performance of MILAMP and is not involved in training. Detailed information about the dataset is available in the supplementary file. It is important to note that annotated hotspot regions of different proteins in this dataset are not precise and may cover an area larger or smaller than the minimal set of amino acids required for amyloid formation.

#### Dataset-4 (DS4)

To analyze the performance of MILAMP for predicting changes in aggregation propensities in amyloid sequences due to point mutations, we have used a dataset of point mutations [15]. This mutation dataset consists of polypeptide sequences together with annotated changes in aggregation propensities due to point mutations. The changes in aggregation propensities (increase, decrease, or no effect) for all 53 mutations in this dataset have been verified experimentally *in vitro* [15]. Detailed information about the dataset is available in the supplementary file.

### 2.2 Feature Extraction

In contrast to existing approaches that typically employ complicated feature extraction techniques, we have used simple amino acid composition features in our model. Specifically, we have used a sliding window approach to calculate the hexapeptide amino acid composition within a protein. This results in a 20-dimensional vector for each hexapeptide within a protein whose components correspond to the normalized frequency of occurrence of different amino acids within a sequence window.

### 2.3 Multiple Instance Learning for Amyloid Prediction

As discussed earlier, there are three major predictive tasks in this domain: 1) classifying amyloid protein from non-amyloid proteins, 2) identifying the subsequences that act as hotspots for amyloid formation and, 3) predicting the change in the aggregation

propensities of a protein due to point mutations. Available datasets for these three problems (DS1-DS4) are structured in such a way that it is not possible for a single classical machine learning model (Support Vector Machines (SVMs), Neural Networks, etc.) to directly generate predictions for all three tasks simultaneously. DS1 and DS2 consists of amino acid sequences which are labelled for amyloidogenicity without any hotspot level annotations. Thus, DS1 can be used to build a classifier for generating hexapeptide-level predictions which can then be used to identify hotspot regions within a protein by employing a sliding window approach. DS2 consists of proteins of varying lengths each with a label indicating whether it can form an amyloid (positive) or not (negative). Consequently, a classical predictive method built using DS2 cannot provide information about the occurrence of hotspots in these proteins and, thus, this dataset cannot be directly used for hotspot prediction. We hypothesize that the combination of DS1 and DS2 can lead to an improved classifier for prediction of amyloid proteins and their hotspots. DS3 provides hotspot level annotations but it is too small (33 proteins only) to have any significant impact on training a hotspot level predictor. Furthermore, DS3 is typically used for benchmarking the predictive accuracy of different machine learning models. DS4 contains information about the effects of mutations on amyloid formation, but it is not possible to directly model amyloid prediction using this dataset with classical machine learning.

To fully exploit available datasets for amyloid prediction, we have employed Multiple Instance Learning (MIL). Multiple instance learning is a form of weak supervision that has been employed in a variety of machine learning problems with ambiguously labeled data. Unlike conventional machine learning problems in which a label is associated with each example, examples in multiple instance learning come in bags [37][38]. A bag is a group of examples and a label is assigned to each bag rather than to individual examples. A positive bag is labeled as positive if it contains at least one positive example as shown in Figure 1. However, it is not known which example in the positive bag is actually positive. If a bag does not contain any positive examples, it is labeled negative. A machine learning model is then built to classify individual instances from such ambiguously labeled data and generate labels at the bag level as well. It is interesting to note that conventional classification is a special case of multiple instance learning with one example per bag [39].

The problem of prediction of hotspots and amyloid forming sequences with annotations at the hexapeptide and protein level in DS1 and DS2, respectively, is ideally suited for multiple instance learning. Specifically, this is achieved by taking each hexapeptide in DS1 and each protein in DS2 as a bag with hexapeptide sub-sequence windows in each protein as examples in the bag. We use protein-level labels (amyloid vs. non-amyloid) as bag labels. We then use a custom machine learning model to learn a classification boundary using this data for classification of hexapeptide windows in a protein as amyloid forming or not. This classification model is then coupled with a ranking model for prediction of mutation effects as discussed in the next two sub-sections.

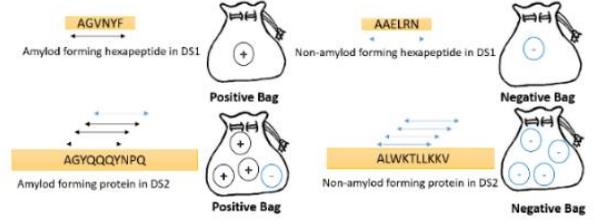


Figure 1 Concept diagram of the MIL formulation for amyloid prediction showing how bags are formed for DS1 and DS2.

### MIL-Classification Model

In our multiple instance learning model, a hexapeptide from DS1 or a protein from DS2 is represented by a bag. A bag corresponding to a DS1 hexapeptide has only one instance whereas a bag corresponding to a DS2 protein has multiple subsequences obtained by overlapping window of hexapeptides. Each of the subsequences is represented by a feature vector  $\mathbf{x}$ . Thus, a single bag corresponding to a hexapeptide (for DS1) or protein (for DS2) can be denoted by  $B_i = \{\mathbf{x}_l, l = 1 \dots n_i\}, i = 1 \dots N$ , where  $n_i$  is the number of hexapeptides in the  $i^{\text{th}}$  protein. The label  $Y_i$  for a given bag indicates whether the corresponding protein is an amyloid (+1) or not (-1). We denote hexapeptide level labels by  $y_l$  which indicate whether a corresponding hexapeptide  $\mathbf{x}_l$  is involved in amyloidogenicity (+1) or not (-1). It is important to note that hexapeptide level labels are available for DS1 only and not DS2. The MIL classification problem can be mathematically expressed as finding a discriminant function  $f(B; \mathbf{w})$ , parameterized by a weight vector  $\mathbf{w}$ , which can predict amyloidogenicity of a given protein represented by bag  $B$ . The prediction score for a given bag can be obtained by taking the maximum linear discriminant score across all examples in the bag. Mathematically, this can be written as:

$$f(B; \mathbf{w}) = \max_{\mathbf{x} \in B} \mathbf{w}^T \mathbf{x}. \quad (1)$$

The MIL classification problem thus requires that  $f(B; \mathbf{w}) > 0$  for positive bags corresponding to amyloid proteins or hexapeptides and  $f(B; \mathbf{w}) < 0$  otherwise. This MIL problem has been solved using our MIL toolbox pyLEMMINGS (PYthon Large Margin Multiple Instance learning System) [27]. PyLEMMINGS models MIL classification as the following optimization problem which is then solved through an iterative stochastic sub-gradient optimization (SSGO) method [28][29].

$$\min_{\mathbf{w}} \rho(\mathbf{B}, \mathbf{Y}; \mathbf{w}) = \frac{\lambda}{2} \|\mathbf{w}\|^2 + \frac{1}{N} \sum_{i=1}^N l(B_i, Y_i; \mathbf{w}) \quad (2)$$

Here  $\|\mathbf{w}\|^2$  is regularization term and  $l(B_i, Y_i; \mathbf{w})$  is the hinge loss function given by:

$$l(B_i, Y_i; \mathbf{w}) = \max\{0, 1 - Y_i f(B_i; \mathbf{w})\} \quad (3)$$

This loss function ensures that the prediction scores  $f(B; \mathbf{w}) = \max_{\mathbf{x} \in B} \mathbf{w}^T \mathbf{x}$  correspond to training labels as discussed above. In SSGO, a bag  $B_{i_t}$  is chosen at an iteration  $t = 1 \dots T$  at random and the objective function with respect to the chosen bag is optimized through a sub-gradient descent step. The objective function at iteration  $t$  can thus be written as:

$$\rho(B_{i_t}, Y_{i_t}; \mathbf{w}_t) = \frac{\lambda}{2} \|\mathbf{w}_t\|^2 + \max\{0, 1 - Y_{i_t} f(B_{i_t}; \mathbf{w}_t)\} \quad (4)$$

In order to perform a weight update, the highest scoring example in the current bag is first identified, i.e.,  $\mathbf{x}_{i_t} = \operatorname{argmax}_{\mathbf{x} \in B_{i_t}} \mathbf{w}_t^T \mathbf{x}$ . Since,  $f(B; \mathbf{w}) = \max_{\mathbf{x} \in B} \mathbf{w}^T \mathbf{x} = \mathbf{w}_t^T \mathbf{x}_{i_t}$ , therefore equation (4) can be written as:

$$\rho(B_{i_t}, Y_{i_t}; \mathbf{w}_t) = \frac{\lambda}{2} \|\mathbf{w}_t\|^2 + \max\{0, 1 - Y_{i_t} \mathbf{w}_t^T \mathbf{x}_{i_t}\} \quad (5)$$

The SSGO solver then performs a descent step using the sub-gradient  $\nabla_t = \partial \rho(B_{i_t}, Y_{i_t}; \mathbf{w}_t) / \partial \mathbf{w}_t$ . Mathematically,

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \nabla_t \quad (6)$$

Here,  $\eta_t = 1/\lambda t$  is the learning rate and  $\nabla_t = \lambda \mathbf{w}_t - \mathbb{I}[Y_{i_t} \mathbf{w}_t^T \mathbf{x}_{i_t} < 1] Y_{i_t} \mathbf{x}_{i_t}$  with  $\mathbb{I}$  denoting the indicator function ( $\mathbb{I}[\cdot] = 1$  iff the argument is true, else 0). The final updated weight vector can be written as follows:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \frac{1}{\lambda t} (\lambda \mathbf{w}_t - \mathbb{I}[Y_{i_t} \mathbf{w}_t^T \mathbf{x}_{i_t} < 1] Y_{i_t} \mathbf{x}_{i_t}) \quad (7)$$

The new weights are used in the objective function for the next iteration. After a fixed number of iterations  $T$ , the weight vector  $\mathbf{w} = \mathbf{w}_{T+1}$  is used in validation and testing. Hotspot prediction for the feature representation of a given hexapeptide can be done by calculating the score  $\mathbf{w}^T \mathbf{x}$  whereas protein level amyloid prediction can be done using  $f(B; \mathbf{w}) = \max_{\mathbf{x} \in B} \mathbf{w}^T \mathbf{x}$ .

### Unified MIL Classification and Ranking Model

In order to produce a unified predictor that can be used to classify amyloid proteins, identify their hotspots and analyze mutation effects, we have modeled the prediction of increase or decrease in amyloidogenicity due to point mutations as an additional ranking constraint in the proposed MIL formulation. For this purpose, we have utilized the mutations dataset (DS4) which consists of a number of point mutations and their experimentally verified effects on amyloid formation. Specifically, we have used features of the wild-type and mutated protein hexapeptide sequences together with the labels indicating the effect of mutations for training. For this purpose, we denote the feature representations of wild-type and mutant subsequences from DS4 by  $\mathbf{x}_j^W$  and  $\mathbf{x}_j^M$ ,  $j = 1 \dots Q$ , respectively.  $y_j^M$  is used to indicate the label for the effect of the mutation (+1 for increased amyloidogenicity and -1 otherwise). The resulting unified MIL problem can be written as the following structural risk minimization:

$$\min_{\mathbf{w}} \rho(\mathbf{w}) = \frac{\lambda}{2} \mathbf{w}^T \mathbf{w} + \beta \sum_I l(B_I, Y_I; \mathbf{w}) + \sum_j r(\mathbf{x}_j^W, \mathbf{x}_j^M, y_j^M; \mathbf{w}) \quad (8)$$

Here,  $\lambda$  is a regularization parameter,  $\beta$  is a scaling parameter that controls the effect of classification (amyloid vs. non-amyloid) errors based on the classification loss in equation-3 and  $r$  is the ranking loss for mutation examples:  $r(\mathbf{x}_j^W, \mathbf{x}_j^M, y_j^M; \mathbf{w}) = \max\{0, 1 - y_j^M \mathbf{w}^T (\mathbf{x}_j^M - \mathbf{x}_j^W)\}$ . (9)

This loss function requires that prediction scores produced by the MIL model correlate with known effects of mutation during training, i.e.,  $\mathbf{w}^T \mathbf{x}_i^M > \mathbf{w}^T \mathbf{x}_i^W$  if amyloid formation increases as a consequence of the mutation and  $\mathbf{w}^T \mathbf{x}_i^M \leq \mathbf{w}^T \mathbf{x}_i^W$  otherwise. This problem is then solved using a stochastic gradient solver in a similar manner as MIL classification through pyLEMMINGS [27]. This formulation allows

us to integrate heterogenous data sources for simultaneous prediction of amyloid proteins, their hotspots and the effects of mutations of amyloid proteins.

## 2.4 Training and Evaluation

In this section, we discuss the training and evaluation protocol used for performance analysis of the proposed method MIL Classification and the Unified MIL Classification and Ranking Models. Area under the Receiver Operating Characteristic Curve (AUC-ROC), expressed in percentage, has been used as a performance metric. In addition to AUC, we have also utilized the area under the receiver operating characteristic curve (AUC-ROC<sub>0.1</sub>) as well as a performance metric to quantify the predictive performance of various methods at low false positive rates. AUC-ROC<sub>0.1</sub> captures the accuracy of the top-scoring predictions from a machine learning model.

### MIL Classification Model

The proposed MIL Classifier (equation 2) is trained on DS1 and evaluated for prediction of amyloid proteins over DS2 using 5-fold cross validation. For cross-validation over DS2, proteins in DS2 are first grouped into 92 clusters using CD-HIT [36] with a sequence identity threshold of 40%. These clusters are then divided into 5 folds each having approximately equal number of label-wise stratified protein sequences. Such a sequence identity-based division into folds ensures that no protein in a given fold shares more than 40% sequence identity with any protein in any other fold. This non-redundancy guarantees that training examples are distinct from testing example across folds and prevents overfitting.

The performance of the proposed model for prediction of hotspot regions and mutation effects is evaluated by using DS3 and DS4 as independent test sets, respectively. However, it is always ensured that for a given test sequence, the training set does not contain any proteins with >40% sequence identity to the test example. For determining the accuracy of hotspot prediction in terms of an ROC curve for a classifier, individual sequence locations in a sequence are first annotated as hotspot (+1) vs. non-hotspot (-1) based on labeling information in DS3. The prediction scores of individual sequence locations are then used to construct the ROC curve across all proteins in DS3.

### Unified MIL Classification and Ranking Model

For the MIL-Rank model (equation 8), DS1 is used for training only whereas 5-fold cross validation is performed over DS2 and DS4 for hotspot and mutation effect prediction, respectively. DS3 is used as an independent test set for evaluating the performance of the predictor over the task of predicting amyloid hotspots. In line with our evaluation protocol for MIL classification, it is always ensured that, for a given test sequence, the training set for this model also does not contain any proteins with >40% sequence identity to the test example.

### External Evaluation

In addition to cross-validation and independent test set analysis over data sets DS1-DS4, we have also performed two external analyses of the proposed scheme which are discussed in detail in the results section.

Firstly, we have analyzed the solvent accessibility of predicted amyloid hotspots in a large non-redundant set of proteins from

the protein data bank. For this purpose, we used a non-redundant dataset of 21,661 protein structures from the protein databank (PDB) [40] and plotted the amyloid forming potential of the top predicted hotspot region within a protein against its relative accessible surface area (rASA) [41]. The hotspot predictions are obtained using the proposed method and rASA is obtained through STRIDE [41]. CD-HIT is used to obtain the set of non-redundant proteins at 40% sequence identity threshold from PDB. We have also analyzed the performance of the proposed scheme over an external set of recently published experimentally verified amyloid proteins from the literature that are not a part of our original datasets. More details on these proteins are given in the results section.

### Hyperparameter Selection and scaling

The hyperparameters of the proposed model ( $\lambda, \beta$ ) have been selected using nested cross-validation over DS-2 using AUC-ROC as the performance metric. Grid search was used to scan the choice of hyperparameters over the range 0.0001 to 1000 in steps of factors of 10.

### 2.6 Code and Webserver

The Python implementation for the proposed method and its webserver are available at the URL:

<http://faculty.pieas.edu.pk/fayyaz/software.html#MILAMP>.

The webserver can be used to obtain amyloid prediction score for a given protein sequence, locate its hotspot region and identify the effects of point mutations. The webserver generates prediction probabilities by scaling raw outputs of our machine learning models through Platt scaling [42].

## 3 RESULTS AND DISCUSSION

The performance the proposed MIL and MIL-Rank predictors has been compared with existing state of the art techniques such as APPNN, MetAmyl and Aggrescan. As a baseline, a simple linear SVM trained over DS1 has been used as well. Table 1 summarizes the results which are discussed below.

### 3.1 Amyloid Prediction

Figure 2 shows the ROC curves of all classifiers for amyloid prediction. The baseline SVM trained on DS1 gives an AUC score of 83.1%. The MIL Classifier trained on DS1 and DS2 using 5-fold cross validation gives a significantly improved AUC score of 88.1% in comparison to the baseline predictor. The MIL-Rank classifier is trained in a similar manner as MIL, but it uses additional information from the mutations dataset DS4 in its training. It gives an AUC score of 85.9%. The performance of MIL and MIL-Rank is compared over DS2 for various existing methods such as APPNN (AUC: 87.9%), MetAmyl (AUC: 88.3%), Aggrescan (AUC: 79.5%) and Waltz (AUC: 71.3%). It is important to note that the performance of the proposed Multiple Instance Learning based approaches is comparable to previous state of the art methods in terms of AUC-ROC. However, the AUC-ROC<sub>0.1</sub> scores of the proposed scheme (53.8% for MIL-RANK) are significantly better in comparison to other schemes (highest score of 44.8% for MetAmyl). Thus, the proposed scheme can be expected to produce fewer false positives in its top predictions.

Table 1: Classification results of different classifiers for amyloid and hotspot prediction. AUC-ROC% scores for Amyloid Prediction over DS2 and Hotspot Prediction over DS3 are reported for each method together with their AUC-ROC<sub>0.1</sub> (in parenthesis). Note that cross-validation (CV) results are reported for DS2 for amyloid prediction whereas DS3 is used for testing only in hotspot prediction. The maximum standard deviation values across multiple runs with different folds for amyloid prediction and hotspot prediction are 1.2% and 0.4%, respectively.

Classifier	Training Data	Amyloid Prediction	Hotspot Prediction
Linear SVM	DS1	83.1 (46.1)	96.8 (76.3)
MIL	DS1+(DS2 CV)	<b>88.1</b> (49.8)	<b>98.0</b> (83.4)
MIL-Rank	DS1+(DS2 CV)	85.9 ( <b>53.8</b> )	97.8 (83.0)
MetAmyl	[19]	<b>88.3</b> (44.8)	96.8 (73.2)
Aggrescan	[10]	79.5 (24.4)	94.1 (66.1)
APPNN	[18]	87.9 (44.2)	97.3 (80.5)

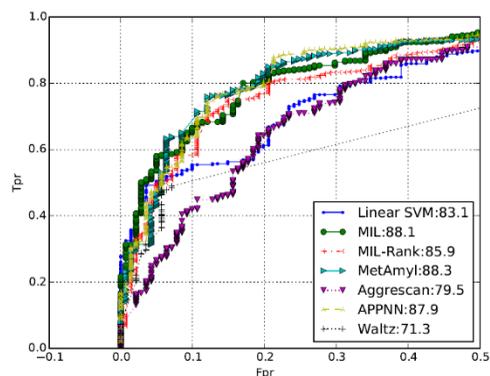


Figure 2 ROC curves for different classification methods for amyloid prediction over DS2. The numbers next to the methods indicate the AUC-ROC in percentage.

### 3.2 Identification of hotspots in polypeptide chains

Table-1 and Figure 3 show the AUC-ROC scores and corresponding ROC curves, respectively, for different hotspot prediction methods. For this purpose, DS3 consisting of 33 proteins with experimentally annotated hotspot regions, has been used as an independent test set. It can be seen that the proposed MIL and MIL-Rank classifiers perform better than existing methods with AUC score of 98% and 97.8%, respectively. This effect is more pronounced at low false positive rates with AUC-ROC<sub>0.1</sub> as the performance metric: MIL gives a score of 83.4% in comparison to 80.5% by APPNN). This shows that the proposed method can be effectively employed to search for hotspot regions within the proteins prior to testing the top candidates in the wet lab. However, it must be pointed out that DS3 is not a complete or precise dataset in that it does not annotate all possible hotspot sequences in proteins and annotated hotspot sequences may cover much larger regions than the set of minimal amino acids required for amyloid formation. However, in the absence of any better datasets and the fact that the same protocol has been used for performance evaluation for all methods, we can expect that the proposed scheme can generalize well over novel test sequences. Identified hotspots for each sequence in DS3 are listed in the supplementary file.

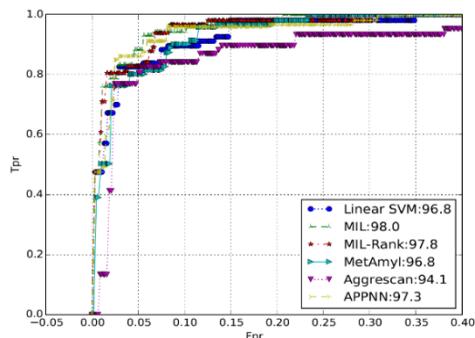


Figure 3 ROC curves for different classification methods for hotspot prediction over DS3. The numbers next to the methods indicate the AUC-ROC in percentage.

### 3.3 Effect of point mutations on amyloidogenicity

Aggrescan is the current state of the art approach for predicting the effect of point mutations and it gives an AUC score of 94.81%. The proposed MIL-Rank method gives an AUC-ROC score of 97.03%. This clearly shows that the proposed scheme is very effective in predicting protein amyloid propensity, hotspots and effects of point mutations in amyloid proteins. Details of individual predictions are given in the supplementary material. Therefore, it can be concluded that the proposed unified model can accurately predict amyloid proteins, their hotspot regions and the effect of mutations on their amyloidogenicity.

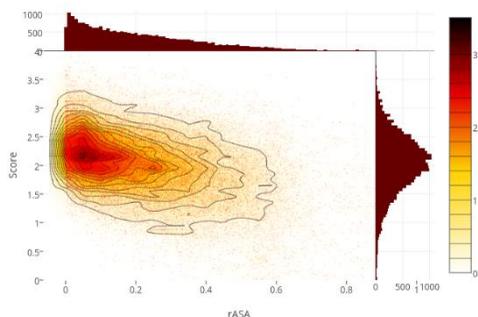


Figure 4 Density plot of amyloid prediction scores vs. relative accessible surface area (rASA) for non-redundant PDB proteins. Each dot represents a predicted amyloid hotspot (with positive raw MILAMP score) in a protein and its rASA within the protein.

### 3.4 Evaluation on non-redundant PDB

Several naturally occurring proteins have amyloid hotspots but are not able to form amyloid fibers. This is because of the fact that the amyloid forming regions in most proteins are not exposed to solvent and occur in the core of the protein [43]. However, amyloid forming proteins contain amyloidogenic segments that are relative surface accessible (rASA) [28]. To evaluate the usability of our method, we have plotted the predicted amyloidogenicity of a potential amyloid forming region within a protein in a non-redundant PDB set against its surface accessibility in Figure 4. It clearly shows that the majority of the high scoring subsequences in proteins with high MILAMP scores have low rASA. This demonstrates that hotspot that promotes amyloidogenesis typically exist inside the proteins and are not able to produce amyloid fibers. Thus, majority of the natural occurring protein do not form amyloid fibers although

they may have an amyloidogenic region. The findings from this large scale analysis are in agreement with the work of Tzotzos and Doig who observed similar patterns over a smaller sample size [43]. It shows that the proposed scheme can be used for studying the behavior of amyloid proteins. It also points out the fact that surface accessibility of hotspot residues must be considered when using hotspot predictions generated from the proposed scheme or from other sequence-based methods as well.

### 3.5 Evaluation on External Proteins

In order to analyze the performance of our proposed model, we have also evaluated it on some recently published experimentally verified amyloid proteins which are not included in any of our datasets (DS1-DS4). The MILAMP webserver was used to generate predictions for various proteins and analyze the concordance of the predictions with experimental findings. We have also used the previous state of the art techniques (Aggrescan and MetAmyl) for this analysis. Below, we discuss the analysis of individual proteins. These results can be easily reconstructed by using the webservers for these methods.

#### TasA protein

Malishev et. al. [44], have experimentally shown that TasA protein (Uniprot id: P54507) interacts with bacterial model membranes which leads to membrane disturbance and structural changes in TasA. TasA forms disordered aggregates which are involved in amyloidogenesis. MILAMP generates amyloid probability score of 0.897) The high prediction score from our model correlates with experimental findings. MetAmyl and Aggrescan also generate positive scores of 0.67 and 0.23, respectively, for this protein. It is interesting to note that the top scoring hotspot regions (135-142 and 52-59) predicted by the proposed scheme are also predicted as hotspots by both MetAmyl and Aggrescan but at much lower ranks. Therefore, we can stipulate that these regions are very good candidates for experimental validation in a future study. MetAmyl and Aggrescan both predict several other hotspot regions as well which are not shared by all three methods.

#### FapC protein

Bleem et al. [45] have examined the sequence of FapC protein (NCBI Reference: WP\_003113480) and experimentally identified specific regions that are involved in amyloid formation. They found three conserved repeats, R1, R2, and R3, each of which contains a GVN<sub>X</sub>A motif (Table 3). The prediction scores of our proposed model for identifying hotspots are also given in Table 2. MILAMP can correctly identify these motifs among its top predictions. This shows the proposed scheme is very effective for predicting hotspot regions in proteins. MetAmyl is also able to identify these regions correctly. However, Aggrescan predicts only one out of the three regions as a hotspot.

#### VL2-8-J1 protein

Brumshtein et. al experimentally identified two segments within the variable domains of Ig light chains using a reference model of VL2-8-J1 (GenBank Id: BAA20021.1) which are involved in forming amyloid fibrils [46]. Each of these segments has been shown to be able to drive amyloid

fibril assembly independently of the other. Thus, these segments are important therapeutic targets. Table 3 shows the amyloid forming regions in the sequence. Our proposed model can correctly identify the second segment as a hot spot as its top prediction. *In contrast, both MetAmyl and Aggrescan predict much large regions as hotspots (70 to 81 and 95-110 by MetAmyl and 73-78 and 97-110 by Aggrescan).*

Table 2: *FapC* protein sequence.

Hotspots	Sequence	Score
R1 (53-101) Motif (83-88)	QQYNNKVSFNGTLNNASVSGSIK- DASGNVGVNVAAGDNNQQANAAALA	0.938
R2 (120-168) Motif (150-155)	QSGYGNTLNNYSNPNTASLS- NSANNVSGNLGVNVAAGFNQKNDLAAA	0.893
R3 (291-324) Motif (307-312)	NNASLSNSLQNVSGNVGVNIAAGGG- NQQSNSLSI	0.917

Table 3: VL2-8-J1 protein sequence with model score.

Amyloid driving segments	Sequence	Model Score	Identified Hotspot
S1 (73-78)	ASLTVS	0.576	ASLTVSG (73-78)
S2 (98-104)	NFYVFGT	0.907	NNFYVFG (97-103)

### HIV-1 Vpu protein

Sneha et. al investigated amyloidogenicity of HIV-1 Vpu protein (Uniprot id: P20882) through molecular dynamics and identified residues 4–35 in the protein to be amyloidogenic [47]. Figure 5 shows that our model produces high score for this region as well. It is interesting to note that the proposed scheme also predicts the same region as an amyloid hotspot. *The same region is also predicted as an amyloid hotspot by both MetAmyl and Aggrescan. However, Aggrescan predicts an additional region of 58–65 as a hotspot.*

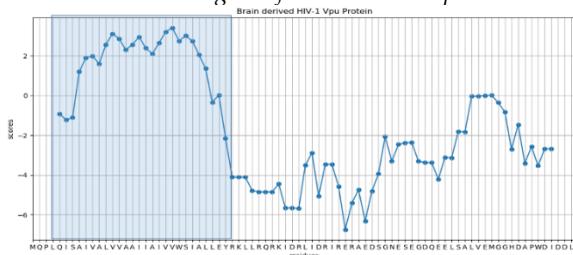


Figure 5 MILAMP raw prediction scores for HIV-1 Vpu protein against its sequence. The known amyloid forming region is highlighted as the shaded region.

## 4 CONCLUSIONS

*In this work, we have proposed a machine learning based method that can simultaneously predict amyloid proteins, their hotspot regions and the effects of point mutations in such proteins. We have shown that the proposed method improves prediction accuracy by integrating heterogenous data sources in modeling the three predictive problems through multiple instance learning. We have also shown that MILAMP can outperform previous state of the art techniques. The proposed scheme can be easily*

*used to generate accurate predictions for a variety of proteins and is expected to prove very useful for studying amyloid proteins.*

## 5 REFERENCES

- [1] M. Belli, M. Ramazzotti, and F. Chiti, "Prediction of amyloid aggregation in vivo," *EMBO Rep.*, vol. 12, no. 7, pp. 657–663, Jul. 2011.
- [2] T. P. J. Knowles, M. Vendruscolo, and C. M. Dobson, "The amyloid state and its association with protein misfolding diseases," *Nat Rev Mol Cell Biol*, vol. 15, no. 6, pp. 384–396, Jun. 2014.
- [3] J. L. Jiménez *et al.*, "Cryo-electron microscopy structure of an SH3 amyloid fibril and model of the molecular packing," *EMBO J.*, vol. 18, no. 4, pp. 815–821, Feb. 1999.
- [4] C. M. Dobson, "Experimental investigation of protein folding and misfolding," *Methods San Diego Calif*, vol. 34, no. 1, pp. 4–14, Sep. 2004.
- [5] J. Greenwald and R. Riek, "Biology of Amyloid: Structure, Function, and Regulation," *Structure*, vol. 18, no. 10, pp. 1244–1260, Oct. 2010.
- [6] C. A. Ross and M. A. Poirier, "Protein aggregation and neurodegenerative disease," *Nat. Med.*, vol. 10 Suppl, pp. S10-17, Jul. 2004.
- [7] F. Chiti and C. M. Dobson, "Protein misfolding, functional amyloid, and human disease," *Annu. Rev. Biochem.*, vol. 75, pp. 333–366, 2006.
- [8] M. I. Ivanova, M. R. Sawaya, M. Gingery, A. Attinger, and D. Eisenberg, "An amyloid-forming segment of beta2-microglobulin suggests a molecular model for the fibril," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 101, no. 29, pp. 10584–10589, Jul. 2004.
- [9] S. Ventura *et al.*, "Short amino acid stretches can mediate amyloid formation in globular proteins: the Src homology 3 (SH3) case," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 101, no. 19, pp. 7258–7263, May 2004.
- [10] C. Nerelius, M. Fitzen, and J. Johansson, "Amino acid sequence determinants and molecular chaperones in amyloid fibril formation," *Biochem. Biophys. Res. Commun.*, vol. 396, no. 1, pp. 2–6, May 2010.
- [11] S. J. Hamodrakas, "Protein aggregation and amyloid fibril formation prediction software from primary sequence: towards controlling the formation of bacterial inclusion bodies," *FEBS J.*, vol. 278, no. 14, pp. 2428–2435, Jul. 2011.
- [12] M. López de la Paz and L. Serrano, "Sequence determinants of amyloid fibril formation," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 101, no. 1, pp. 87–92, Jan. 2004.
- [13] R. Zambrano, M. Jamroz, A. Szczasiuk, J. Pujols, S. Kmiecik, and S. Ventura, "AGGRESAN3D (A3D): server for prediction of aggregation properties of protein structures," *Nucleic Acids Res.*, vol. 43, no. W1, pp. W306–W313, Jul. 2015.
- [14] K. Sankar, S. R. Krystek, S. M. Carl, T. Day, and J. K. X. Maier, "AggScore: Prediction of aggregation-prone regions in proteins based on the distribution of surface patches," *Proteins*, vol. 86, no. 11, pp. 1147–1156, 2018.
- [15] O. Conchillo-Solé, N. S. de Groot, F. X. Avilés, J. Vendrell, X. Daura, and S. Ventura, "AGGRESAN: a server for the prediction and evaluation of 'hot spots' of aggregation in polypeptides," *BMC Bioinformatics*, vol. 8, no. 1, p. 65, Feb. 2007.
- [16] N. Sánchez de Groot, I. Pallarés, F. X. Avilés, J. Vendrell, and S. Ventura, "Prediction of 'hot spots' of aggregation in disease-linked polypeptides," *BMC Struct. Biol.*, vol. 5, p. 18, Sep. 2005.
- [17] S. O. Garbuzynskiy, M. Y. Lobanov, and O. V. Galzitskaya, "FoldAmyloid: a method of prediction of amyloidogenic regions from protein sequence," *Bioinforma. Oxf. Engl.*, vol. 26, no. 3, pp. 326–332, Feb. 2010.
- [18] A. P. Pawar, K. F. DuBay, J. Zurdo, F. Chiti, M. Vendruscolo, and C. M. Dobson, "Prediction of 'Aggregation-prone' and 'Aggregation-susceptible' Regions in Proteins Associated with Neurodegenerative Diseases," *J. Mol. Biol.*, vol. 350, no. 2, pp. 379–392, Jul. 2005.
- [19] G. G. Tartaglia, A. P. Pawar, S. Campioni, C. M. Dobson, F. Chiti, and M. Vendruscolo, "Prediction of aggregation-prone regions in structured proteins," *J. Mol. Biol.*, vol. 380, no. 2, pp. 425–436, Jul. 2008.
- [20] G. G. Tartaglia and M. Vendruscolo, "The Zyggregator method for predicting protein aggregation propensities," *Chem. Soc. Rev.*, vol. 37, no. 7, pp. 1395–1401, Jul. 2008.
- [21] A.-M. Fernandez-Escamilla, F. Rousseau, J. Schymkowitz, and L. Serrano, "Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins," *Nat. Biotechnol.*, vol. 22, no. 10, pp. 1302–1306, Oct. 2004.
- [22] S. Maurer-Stroh *et al.*, "Exploring the sequence determinants of amyloid structure using position-specific scoring matrices," *Nat. Methods*, vol. 7, no. 3, pp. 237–242, Mar. 2010.

- [23] M. Oliveberg, "Waltz, an exciting new move in amyloid prediction," *Nat. Methods*, vol. 7, no. 3, pp. 187–188, Mar. 2010.
- [24] C. Família, S. R. Dennison, A. Quintas, and D. A. Phoenix, "Prediction of Peptide and Protein Propensity for Amyloid Formation," *PLOS ONE*, vol. 10, no. 8, p. e0134679, Aug. 2015.
- [25] M. Burdukiewicz, P. Sobczyk, S. Rödiger, A. Duda-Madej, P. Mackiewicz, and M. Kotulska, "Amyloidogenic motifs revealed by n-gram analysis," *Sci. Rep.*, vol. 7, no. 1, p. 12961, 11 2017.
- [26] P. Gasior and M. Kotulska, "FISH Amyloid - a new method for finding amyloidogenic segments in proteins based on site specific co-occurrence of aminoacids," *BMC Bioinformatics*, vol. 15, p. 54, Feb. 2014.
- [27] M. Emily, A. Talvas, and C. Delamarche, "MetAmyl: a METa-predictor for AMYLOid proteins," *PloS One*, vol. 8, no. 11, p. e79722, 2013.
- [28] K. K. Frousios, V. A. Iconomidou, C.-M. Karletidi, and S. J. Hamodrakas, "Amyloidogenic determinants are usually not buried," *BMC Struct. Biol.*, vol. 9, p. 44, Jul. 2009.
- [29] A. C. Tsolis, N. C. Papandreou, V. A. Iconomidou, and S. J. Hamodrakas, "A consensus method for the prediction of 'aggregation-prone' peptides in globular proteins," *PloS One*, vol. 8, no. 1, p. e54175, 2013.
- [30] J. Beerten *et al.*, "WALTZ-DB: a benchmark database of amyloidogenic hexapeptides," *Bioinforma. Oxf. Engl.*, vol. 31, no. 10, pp. 1698–1700, May 2015.
- [31] P. P. Wozniak and M. Kotulska, "AmyLoad: website dedicated to amyloidogenic protein fragments," *Bioinforma. Oxf. Engl.*, vol. 31, no. 20, pp. 3395–3397, Oct. 2015.
- [32] D. M. Rowczenio *et al.*, "Online registry for mutations in hereditary amyloidosis including nomenclature recommendations," *Hum. Mutat.*, vol. 35, no. 9, pp. E2403-2412, Sep. 2014.
- [33] G. De Baets *et al.*, "SNPeffect 4.0: on-line prediction of molecular and structural effects of protein-coding variants," *Nucleic Acids Res.*, vol. 40, no. Database issue, pp. D935-939, Jan. 2012.
- [34] M. J. Landrum *et al.*, "ClinVar: public archive of interpretations of clinically relevant variants," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D862-868, Jan. 2016.
- [35] F. ul A. Afsar Minhas, E. D. Ross, and A. Ben-Hur, "Amino acid composition predicts prion activity," *PLoS Comput. Biol.*, vol. 13, no. 4, Apr. 2017.
- [36] W. Li and A. Godzik, "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences," *Bioinforma. Oxf. Engl.*, vol. 22, no. 13, pp. 1658–1659, Jul. 2006.
- [37] S. Andrews, T. Hofmann, and I. Tsochantaridis, "Multiple Instance Learning with Generalized Support Vector Machines," in *Eighteenth National Conference on Artificial Intelligence*, Menlo Park, CA, USA, 2002, pp. 943–944.
- [38] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support Vector Machines for Multiple-instance Learning," in *Proceedings of the 15th International Conference on Neural Information Processing Systems*, Cambridge, MA, USA, 2002, pp. 577–584.
- [39] E. Alpaydin, *Introduction to Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2004.
- [40] S. K. Burley *et al.*, "RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D464–D474, Jan. 2019.
- [41] M. Heinig and D. Frishman, "STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins," *Nucleic Acids Res.*, vol. 32, no. Web Server issue, pp. W500-502, Jul. 2004.
- [42] J. C. Platt, "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods," in *Advances in Large Margin Classifiers*, 1999, pp. 61–74.
- [43] S. Tzotzos and A. J. Doig, "Amyloidogenic sequences in native protein structures," *Protein Sci. Publ. Protein Soc.*, vol. 19, no. 2, pp. 327–348, Feb. 2010.
- [44] R. Malishev, R. Abbasi, R. Jelinek, and L. Chai, "Bacterial Model Membranes Reshape Fibrillation of a Functional Amyloid Protein," *Biochemistry*, vol. 57, no. 35, pp. 5230–5238, 04 2018.
- [45] A. Bleem *et al.*, "Protein Engineering Reveals Mechanisms of Functional Amyloid Formation in *Pseudomonas aeruginosa* Biofilms," *J. Mol. Biol.*, vol. 430, no. 20, pp. 3751–3763, Oct. 2018.
- [46] B. Brumshtein, S. Esswein, M. R. Sawaya, A. T. Ly, M. Landau, and D. S. Eisenberg, "Identification of two principal amyloid-driving segments in variable domains of Ig light chains in AL amyloidosis," *bioRxiv*, p. 354571, Jun. 2018.
- [47] P. Sneha, P. K. Panda, F. R. Gharemirshamlu, K. Bamdad, and S. Balaji, "Structural discordance in HIV-1 Vpu from brain isolate alarms amyloid fibril forming behavior- a computational perspective," *J. Theor. Biol.*, vol. 451, pp. 35–45, Aug. 2018.

**Farzeen Munir** received BS and MS degrees in Electrical and Systems Engineering from Pakistan Institute of Engineering and Applied Sciences, Pakistan. She is currently pursuing her PhD degree at Gwangju Institute of Science and Technology, Korea in Electrical Engineering and Computer Science. Her current research interest includes machine Learning, deep neural network and computer vision.

**Sadaf Gull** holds a degree in BS Computer Science from University of Punjab, Pakistan and an MS Computer Science degree from the Government College University, Lahore, Pakistan. Currently she is working towards her PhD degree in Computer Science from PIEAS and is working on machine learning based protein function annotation. Her Ph.D. is supported by a grant from the indigenous Ph.D. fellowship scheme of the Higher Education Commission (HEC) of Pakistan.

**Amina Asif** holds an MS degree in Computer Science from the Pakistan Institute of Engineering and Applied Sciences. Currently she is working on the development of weak-supervision techniques in machine learning as part of her PhD at Pakistan Institute of Engineering and Applied Sciences. She is funded through IT and Telecom Endowment Fund at PIEAS.

**Fayyaz Minhas** holds an MS degree in Systems Engineering from Pakistan Institute of Engineering and Applied Sciences. Dr. Minhas is a recipient of Fulbright Scholarship for PhD Computer Science studies at Colorado State University in the area of machine learning in bioinformatics. He is currently the principal investigator for Biomedical Informatics and Data Science labs in the Department of Computer and Information Sciences, PIEAS. His complete profile is available at: <http://faculty.pieas.edu.pk/fayyaz/>.